

DOCUMENT RESUME

ED 194 535

TM 800 663

AUTHOR Scheetz, James P.; vonFraunhofer, J. Anthony
TITLE Measuring Criterion-Referenced Test Reliability with
a Single Test Administration.
PUB DATE Apr 80
NOTE 19p.: Paper presented at the Annual Meeting of the
National Council on Measurement in Education (Boston,
April 7-11, 1980).
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Cutting Scores; Error of
Measurement; Higher Education; *Mastery Tests;
*Mathematical Models; Multiple Choice Tests; *Test
Reliability
IDENTIFIERS *Subkoviak (Michael J)

ABSTRACT

Subkoviak suggested a technique for estimating both group reliability and the reliability associated with assigning a given individual to a mastery or non-mastery category based on a single test administration. Two assumptions underlie this model. First, it is assumed that had successive test administrations occurred, scores for each individual would be independent over repeated testings. Second, it is assumed that scores for each individual are binomially distributed. According to Subkoviak, these assumptions are frequently violated. The purpose of this investigation was to compare the results obtained from his model, for both a simple and compound binomial distribution of test scores, with classical estimates of reliability: Guttman, and KR-20 and KR-21. An inverse relationship existed between the two types of estimates. For students who exhibited no learning on successive test administrations, a comparison of the reliability estimates indicates that the Subkoviak estimates are less sensitive to a small observed score variance. However, the model provides useful reliability estimates in situations where the classical measures are clearly inappropriate. The model gave reasonable reliability estimates even though its assumptions were not met. (Author/CP)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED194535

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.



Measuring Criterion-Referenced Test
Reliability with a Single Test
Administration

James P. Scheetz, Ph.D.
Assistant Professor of
Community Dentistry

J. Anthony vonFraunhofer, Ph.D.
Professor of
Biomaterials Science

University of Louisville
Health Sciences Center
P.O. Box 35260
Louisville, Kentucky 40292

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. P. Scheetz

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Presented at the annual meeting of the
National Council on Measurement in Education
April 8-10, 1980
Boston, Massachusetts

TM 800 663

Abstract

The Subkoviak method for estimating test reliability from a single administration of a mastery test was used in a situation in which the assumptions of the model were violated. A 100 item multiple choice test was administered four times to a group of 33 students. The Subkoviak reliability estimates were compared with classical reliability estimates and an inverse relationship was found to exist between the two types of estimates. For those groups of students who exhibited no learning on successive test administrations, a comparison of the reliability estimates indicates that the Subkoviak estimates are less sensitive to a small observed score variance. However, the Subkoviak method provides useful reliability estimates in situations where the classical measures are clearly inappropriate. The findings from this study support the use of the Subkoviak model for estimating reliability when mastery tests are used to assess student achievement.

In a recent publication, Subkoviak (1976) suggested a technique for estimating both group reliability and the reliability associated with assigning a given individual to a mastery or non-mastery category based on a single test administration. Two assumptions underlie this model. First, it is assumed that had successive test administrations occurred, the scores achieved by each individual would be independent over repeated testings. Second, it is assumed that scores for each individual are binomially distributed. As Subkoviak (1976) has pointed out, these assumptions are frequently violated. The purpose of this investigation was to compare the results obtained from the Subkoviak model, for both a simple and compound binomial distribution of test scores, with classical estimates of reliability which are commonly obtained when performing item analysis procedures.

Algina and Noe (1978) investigated the accuracy of Subkoviak's approach using two true score estimates: (1) an estimate obtained by linear regression and (2) the observed proportion-correct score. The regression estimate resulted in underestimates of the reliability for cutting scores near the test mean and overestimates of the reliability when the cut-off scores were relatively far from the test mean. When the proportion-correct estimator was used, overestimates of the reliability were obtained for cut-off scores near the test mean and underestimates were obtained for cutting scores closer to extreme values. Subkoviak (1978) concluded that the compound binomial procedure produces reliability estimates which are too small for cut-off scores near the mean of the test scores and too large for cut-off scores in the tails of the test score distribution for short tests of ten items or less. However, the

estimation procedure generally provides accurate reliability estimates.

Specifications of the Model

The mathematical model given by Subkoviak permits the calculation of a coefficient of agreement for each individual. This is defined as the probability that a given individual is assigned to the same mastery state as a result of responding to parallel tests X and X' . For two mastery levels (mastery and non-mastery), the probability of consistent classification of any individual, $P_c^{(i)}$, is given by

$$P_c^{(i)} = \text{Prob}(X_i \geq C, X'_i \geq C) + \text{Prob}(X_i < C, X'_i < C) \quad (1)$$

where

- C = cut-off score for determining mastery
- X_i = observed score for examinee i on test X
- X'_i = observed score for examinee i on test X'

By invoking the assumption that the scores on X and X' are independently distributed for each person and that the scores on the two tests are identically binomially distributed, Subkoviak showed that Equation (1) can be rewritten as:

$$P_c^{(i)} = [\text{Prob}(X_i \geq C)]^2 + [\text{Prob}(X_i < C)]^2 \\ = [\text{Prob}(X_i \geq C)]^2 + [1 - \text{Prob}(X_i \geq C)]^2 \quad (2)$$

where

$$\text{Prob } (X_i \geq C) = \sum_{X_i=C}^n \binom{n}{X_i} P_i^{X_i} (1 - P_i)^{n - X_i} \quad (3)$$

In Equation (3), n is the number of items on the test and P_i is the probability of a correct item response. The reader is referred to Subkoviak (undated) for equations which can be used to compute $P_c^{(i)}$ with multiple cut-off scores.

Using Equations (2) and (3), the coefficient of agreement can be computed for each individual. In addition to a coefficient for each person, Subkoviak has defined the group coefficient of agreement as the mean of the individual coefficients. This group coefficient is given by

$$P_c = \frac{\sum_{i=1}^N P_c^{(i)}}{N} \quad (4)$$

where N is the total number of examinees.

The Testing Procedure

The data analyzed in this study was obtained from multiple testings of dental students in a required dental management course. The test consisted of 19 sections, comprising overall 100 multiple choice items, each having four options. Every test item was specifically related to objectives which the students were given at the beginning of the course.

The testing procedure was designed to assess mastery which was defined as achieving a passing grade in the course. A passing grade was established as a score of 70 or above. However, in the overall scoring sys-

tem used in the dental school, it was required that letter grades be assigned; these grades were assigned on the following basis: A = 90 - 100, B = 80 - 89, C = 70 - 79, F, E ≤ 69. Therefore, a student could be assigned to one of four mastery levels with the cutting points being 90, 80, and 70. Since previous experience indicated that the examination is difficult and only one-third of the students could be expected to achieve a passing grade on the first administration of the exam, multiple administrations were instituted.

The students were required to take the test each time it was offered until they received a score of 70 or above. After receiving a passing grade, students then had the option of taking the exam additional times in an attempt to raise their grade with the provision that the highest grade achieved on any testing would be the final course grade. Each student was permitted to take the exam a maximum of five times.

After every testing, each student was given a computer print-out summarizing his or her performance and indicating how many exam questions had been answered correctly. In addition, the students were told how many questions in each section were answered correctly. This mechanism was used to inform the students which sections they should review to prepare for subsequent testings.

Method

The 100 item test was divided into two 50 item subtests by assigning the even items to one subtest and the odd items to a second subtest. The means and standard deviations were calculated and the coefficient of equivalence (reliability) computed by a method proposed by Guttman (1945).

In this, equal variability of the two subtests is not assumed and the formula used to compute the reliability is

$$r_{XX} = 2[1 - \frac{S^2_E + S^2_0}{S^2_X}] \quad (5)$$

where S^2_E and S^2_0 are the variances of the two subtests and S^2_X is the total score variance.

The reliability estimate obtained from Equation (5) was compared with Subkoviak's group coefficient of agreement where the true score for each examinee was estimated by a regression procedure (See Subkoviak, 1976, p. 269). In addition to the Guttman reliability estimate, KR-20 and KR-21 were computed since these are commonly obtained reliability estimates. It is recognized, however, that KR-20 and KR-21 may be inappropriate for use with criterion-referenced tests.

In addition to comparing the results obtained from all examinees ($N = 33$), a comparison was made between those students who exhibited no learning on successive administrations of the test. No learning was defined as receiving the same letter grade on successive testings. Twenty students showed "no learning" from the first to the second testing. From the second to the third testing, ten students fell into the "no learning" category; and nine students showed "no learning" from the third to the fourth testing.

It would be expected that the Subkoviak reliability estimates would closely agree with those obtained from Equation (5) since the "no learning" condition could be viewed as satisfying the assumption that test scores on successive administrations are independent.

Results

In this section, the results will be presented for both the classical procedures and the Subkoviak procedures and the results of the two procedures will be compared.

Classical Approach

The results over the four testings of all examinees are shown in Table 1. For the first testing, the KR-20 estimate exceeds both KR-21 and the Guttman estimate. For the second administration, the same relationship which was found for the first testing exists among the three estimates although the amount by which the KR-20 value exceeds the other two estimates is smaller. The third testing resulted in almost identical values for KR-20 and the Guttman estimate, both of which are greater than the KR-21 value. The fourth testing resulted in a decrease of all three estimates when compared with the estimates obtained from the third testing. The largest decrease occurred with KR-21 and the smallest decrease was obtained with KR-20. Of the three estimates, the KR-20 estimates appear to change the least over repeated testings, whereas the Guttman estimates fluctuate the most.

For those testings which included only those students who exhibited "no learning" on successive test administrations, the classical reliability estimates are quite variable. These results are reported in Tables 2, 3 and 4. A comparison of the two test administrations reported in Table 2 indicates that the KR-20 values exceed both the KR-21 and Guttman estimates. The same result occurred when the values from the second and third administrations (Table 3) were compared; the KR-20 estimates are greater

than the other two classical reliability estimates. A comparison of the third and fourth testings (Table 4) shows that both KR-20 and Kr-21 are 0.0 and the Guttman estimate is 0.04 for the third testing. On the fourth administration, the KR-21 estimate is much less than the Guttman and KR-20 estimates.

Subkoviak Approach

The results obtained from the Subkoviak procedure over all examinees for the four testings (Table 1) show that the compound binomial reliability estimates exceed the simple binomial estimates by 0.05 on the first and second testings, by 0.06 for the third administration, and by 0.08 on the fourth testing. These results are consistent with the findings presented by Subkoviak (1976) in which the compound binomial estimates exceeded the simple binomial estimates.

For those test administrations which involved no learning by the students, the compound binomial estimates exceeded the simple binomial estimates in all instances. A comparison of the first and second administrations (Table 2) shows that the compound estimates exceed the simple estimates by 0.05. The difference between the two estimates was 0.05 and 0.04 respectively, when comparing the second and third testings (Table 3). A comparison of the third and fourth administrations (Table 4) shows that the differences are 0.05 and 0.07 respectively.

Comparison of Classical and Subkoviak Procedures

For purposes of comparing the classical approach with the Subkoviak procedure, the Guttman reliability estimates will be compared with the

compound binomial reliability estimates. The reason for choosing these two estimates is that the Guttman estimate does not assume that the variances on the two halves of a test are equal, whereas the use of KR-20 and KR-21 makes this assumption. These latter two estimates were reported because they are frequently obtained when item analysis procedures are carried out. The compound binomial estimate was chosen over the simple binomial estimate because the compound estimate..."provides an even better approximation to reality in most testing situations," (see Subkoviak, 1976, p. 268).

The results from the four testings of all students indicate that an inverse relationship exists between the Guttman and compound reliability estimates. The Guttman estimate increased from the first to the second, and the second to the third testing; and decreased from the third to the fourth testing. The compound binomial estimates decreased from the first to the second, and the second to the third administrations; but increased from the third to the fourth testing.

For those testings based on a comparison of those students who exhibited no learning, the compound reliability estimates exceeded the Guttman estimates when comparing the first and second testings (Table 2). This relationship was reversed when comparing the second and third administrations (Table 3). The results reported in Table 4 show that the compound estimate greatly exceeds the Guttman estimate for the third testing. The fourth testing shows this same relationship although the difference between the two estimates is not as great.

Discussion

Several findings from this study are of interest. First, the Subkoviak model does not appear to be sensitive to small observed test score variance. While it is recognized that classical reliability estimates and the Subkoviak estimates may both be low when the test score variance is small, the Subkoviak estimates did not decrease when this occurred. This conclusion is most evident from the results presented in Table 4. This finding suggests that the binomial methods developed by Subkoviak appear to result in reliability estimates which are not dependent upon maximizing score variations among individuals.

The second finding of interest is that the Guttman estimates appear to vary inversely with both binomial estimates for those testings involving all students. This finding may be explained by Subkoviak's contention that if, on successive testings the results are positively correlated, the single administration reliability will be a conservative estimate of actual reliability. Conversely, if the dual administration results are negatively correlated, the single administration estimate will be liberal when compared with the Guttman estimate.

The third and last point of interest is that the Subkoviak reliability estimates appear to be sensitive to the cut-off score points. For the four testings involving all examinees, the Subkoviak estimates are greater for the first testing than for any succeeding testing. A possible explanation for this finding is that the test mean is six points below the nearest cut-off score for the first testing, but less than one-half point from a cut-off score on succeeding administrations. It is suggested that the selection of cut-off score points will influence the

simple and compound reliability estimates obtained when using the Subkoviak procedures.

There are several limitations inherent in the present study. First, it was not possible to assess the potential bias which may result from the estimation procedures used in this investigation since the findings were based on classroom rather than simulated data. It was not possible to compute standard deviations for replications since there were no replications per se. A second limitation is the relatively small number of examinees involved in some testings. As Algina and Noe (1978) have noted, the variability of the estimates is expected to increase as the number of examinees decreases. However, they also reported that the bias is not affected by varying the number of examinees. A third limitation is the test length of 100 items. This number is much greater than the number of items typically used to assess mastery. Due to scheduling constraints, it was not possible to use a series of shorter tests. In order to examine the students over the entire range of content, a 100 item test was developed and used by the course faculty.

Conclusion

The purpose of this investigation was to assess the usefulness of the Subkoviak model for estimating the reliability of criterion-referenced tests when the assumptions of the model are violated. The findings from this study suggest that the Subkoviak procedure results in reasonable reliability estimates when the assumptions are not met. Further, classical reliability estimates are clearly inappropriate when the test score variance is small. Therefore, the Subkoviak model appears to be a viable method for obtaining reliability estimates of criterion-referenced tests.

References

- Algina, J. and Noe, M. J. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 1978, 15, 101-110.
- Guttman, L. A basis for analysing test-retest reliability. Psychometrika, 1945, 10, 255-282.
- Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.
- Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.
- Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. (Occasional Paper No. 15), Madison: University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design, undated.

Table 1
Means, Standard Deviations and Reliabilities of Four
Test Administrations
N = 33

	FIRST ADMINISTRATION			SECOND ADMINISTRATION			THIRD ADMINISTRATION			FOURTH ADMINISTRATION		
	All Items	Even Items	Odd Items	All Items	Even Items	Odd Items	All Items	Even Items	Odd Items	All Items	Even Items	Odd Items
Mean	63.91	30.73	33.18	70.10	33.55	36.55	80.37	38.73	41.64	90.18	45.42	44.76
Standard Deviation	7.50	4.19	4.74	8.24	5.19	4.18	6.98	4.71	2.83	4.61	2.97	2.31
Simple Binomial Reliability	.78			.62			.59			.65		
Compound Binomial Reliability	.83			.67			.65			.73		
Guttman Reliability	.58			.69			.76			.67		
KR-20	.72			.77			.77			.74		
KR-21	.60			.70			.68			.59		

Table 2
Means, Standard Deviations and Reliabilities
for No Learning on First and Second
Administrations
N = 20

	FIRST ADMINISTRATION			SECOND ADMINISTRATION		
	All Items	Even Items	Odd Items	All Items	Even Items	Odd Items
Mean	63.10	30.55	32.55	66.75	31.75	35.00
Standard Deviation	8.47	4.45	5.24	8.34	5.28	4.13
Simple Binomial Reliability	.81			.70		
Compound Binomial Reliability	.86			.75		
Guttman Reliability	.68			.70		
KR-20	.79			.78		
KR-21	.68			.69		

Table 3
Means, Standard Deviations and Reliabilities
for No Learning on Second and Third
Administrations
N = 10

	SECOND ADMINISTRATION			THIRD ADMINISTRATION		
	All Items	Even Items	Odd Items	All Items	Even Items	Odd Items
Mean	75.60	36.90	38.70	76.90	36.40	40.50
Standard Deviation	9.07	5.15	4.81	9.69	6.29	4.17
Simple Binomial Reliability	.60			.58		
Compound Binomial Reliability	.65			.62		
Guttman Reliability	.80			.78		
KR-20	.84			.87		
KR-21	.78			.82		

Table 4
Means, Standard Deviations and Reliabilities
for No Learning on Third and Fourth
Administrations
N = 9

	THIRD ADMINISTRATION			FOURTH ADMINISTRATION		
	All Items	Even Items	Odd Items	All Items	Even Items	Odd Items
Mean	81.67	39.89	41.78	85.23	42.67	42.56
Standard Deviation	2.69	1.62	2.73	4.41	3.12	2.13
Simple Binomial Reliability	.58			.67		
Compound Binomial Reliability	.63			.70		
Guttman Reliability	.04			.54		
KR-20	.00			.57		
KR-21	.00			.36		